DATA PRIVACY IN WEB ANALYTICS

An Empirical Study and Declaration Model of Data Collection on Websites

Darius Zumstein, Aleksandar Drobnjak and Andreas Meier
Information Systems Research Group, University of Fribourg, Boulevard de Pérolles 90, 1700 Fribourg, Switzerland

Keywords: Web analytics, Web controlling, Google Analytics, Web analytics systems, Data collection, Data privacy, Privacy policy, Transparency, Declaration.

Abstract: Web analytics has become a useful instrument for electronic business and website management to analyze and optimize website usage. However, different concerns arise in web analytics regarding the collection, storage and usage of web data. To maintain user confidence in websites, operators need to comply with privacy and to inform truthfully about data collection. One way to achieve transparency would be by means of informing users on the purpose, methods and processes of data collection taking place and on the subsequent analysis and use of such. Results of an empirical study conducted show that 16% of the Forbes 500 listed companies, do not declare their data collection practices at all. Moreover, 35% do not declare the usage of cookies and 61% do not declare the recording of IP addresses. Surprisingly, 91% of the websites do not name the web analytics system (provider) used to track web usage and user data. A large backlog exists regarding transparency of data collection in privacy policies, especially when Google Analytics is used. This paper demands more transparency in web analytics and proposes a declaration model with seven criteria for the evaluation of data collection on websites: type of data being collected, purpose of data collection, method of data collection as well as the technology and software used for data collection. In addition, websites should provide the right to access personal data and to deactivate data collection during visits.

1 INTRODUCTION

In the digital information and Internet age, websites have become a crucial instrument of information, communication and business transaction. Therefore, *web analytics*, which is defined as the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing web usage (WAA, 2011), is now an important issue for both business practice and academic research. According to a study by Forrester Research, 74% of large enterprises state that web analytics is a technology they cannot do without (Forrester, 2009).

Web analytics allows website managers to optimize the website (navigation, content or usability), online marketing (e.g. campaigns or search engine optimization) and customer relationship management, i.e. customer orientation, acquisition and retention (Kaushik, 2009, Meier and Zumstein, 2010).

Although web analytics has many benefits, some problems exist. The most critical and publicly discussed problem in web analytics is *data privacy*. Web analytics systems are capable of tracking every visitor's click. If these web usage data is linked to personal data or user profiles, privacy issues become highly relevant for web analytics. Nevertheless, little research in web information systems has focused on privacy in web analytics. This paper seeks to answer the central research question, how transparently corporate websites inform their users on data collection. In doing so, following sub-questions are posed:

1) How many companies collect data on their website – and what kind of data is being collected?
2) How – with what methods, technologies and software – is data collected on websites?
3) Do website operators correctly and transparently declare data collection in their privacy policies?
4) What are possible criteria, best practices and recommendations regarding transparent declaration of data collection on websites?

After the problem statement in section 2, section 3 summarizes the results of the study of declaration of data collection on websites. Section 4 proposes a declaration model of data collection. Section 5 con

Zumstein D., Drobnjak A. and Meier A..
DATA PRIVACY IN WEB ANALYTICS - An Empirical Study and Declaration Model of Data Collection on Websites.
DOI: 10.5220/0003301404660472
In *Proceedings of the 7th International Conference on Web Information Systems and Technologies* (WEBIST-2011), pages 466-472
ISBN: 978-989-8425-51-5

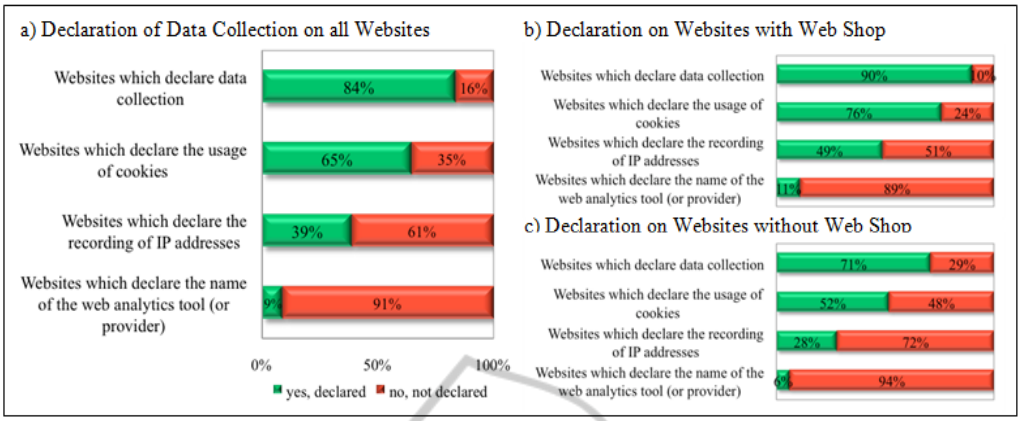

Figure 1: Declaration of data collection on the Forbes Global Top 500 companies' websites (n = 500).

cludes with recommendations to visitors, data protection officers and webmasters and with an outlook.

2 PROBLEM STATEMENT

Similar to hospitals or banks, where personal data of patients or clients is collected, privacy of Internet users is involved, if cookies, IP addresses or login data are used to identify users and if *web analytics tools* track the interactions of visitors with websites. User data – like preferred web pages, history of click behaviour, location of access, browser settings and access data – has to be protected too. Data privacy becomes especially critical, if web data and metrics (including page views, visits or conversion rates) are linked to sales data or to personal data such as name, e-mail, postal address or credit card number.

Most countries have passed privacy or data protection acts in order to guarantee correct and confidential handling of personal data. Nevertheless, privacy regulations differ from country to country, and privacy policies, terms or disclaimers also vary from website to website.

However, if privacy policies do not conform with national or international privacy acts, if they are vaguely formulated, and if declarations regarding data collection are imprecise, incomplete, incorrect or missing, *trust and positivity* of visitors towards websites can be disturbed and personal interactivity with websites or companies negatively influenced.

In contrast, clearly defined privacy policies and transparent declaration of data collections increase the *trustworthiness*, *reputation and attractiveness* of websites and organizations. In light of this, privacy policies and declaration should not be understood as regulatory barriers, but rather as a bridge between visitors and websites. They are basics to establish long-term, profitable relationships between visitors and website operators, and between online customers and (electronic business) companies.

3 RESULTS OF THE STUDY

3.1 Research Method

To study data collection and declaration on websites, the top 500 companies of the *Forbes Global 2000* were selected (Forbes, 2011). Besides reading and personally analyzing the privacy policies and declarations of all 500 corporate websites, the tool *Web Analytics Solution Profiler* (WASP, 2011) was used to identify the *web analytics system(s)* used on every website. WASP is available as a plug-in for Mozilla Firefox and identifies more than 200 web analytics tools, using *client-side data collection methods* (page tagging). Server-side data collection methods (logfile analysis) or other methods, like reverse proxies or packet sniffing, cannot be analyzed by WASP and are not considered in this study.

The privacy policies, terms and disclaimers of the Forbes Global 500 corporate websites were analyzed in June and July 2010 with regard to

- the declaration of the usage of *cookies*
- the collection of *IP addresses*
- the mention of the name of a *web analytics tool*.

3.2 Results on General Declaration

The analysis of the privacy policies on the 500 websites resulted in the following conclusions: 305 (or 84%) of the 365 websites, which are using web analytics, vaguely mention in their privacy policies

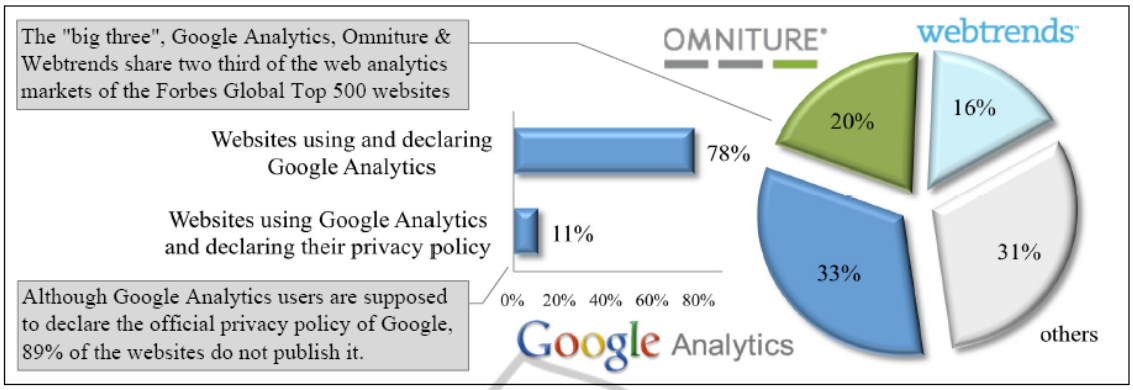

Figure 2: Market share and declaration of Google Analytics on the Forbes Global Top 500 companies' websites (n = 500).

that data is collected on the website: *16%* of the firms *do collect data* but *do not declare* this at all, what is legally problematic (Figure 1a).

Only 65% of the companies declare that cookies are used to identify returning visitors.

This means, on *every third* website the user is *not informed* that a cookie was set on his computer. This result is alarming, as most European countries mandate that the usage of cookies is declared on the site.

Moreover, only 39% of the 500 websites state that they collect Internet Protocol (IP) addresses of visitors. This means that nearly *two thirds do not declare the collection of IP addresses*, even though the IP address can be considered personal data. In addition, only 9% of the websites explicitly mention the name of the web analytics software used, or the name of the tool provider. In other words: *91% do use web analytics systems* to analyze website usage, but *do not declare* which tool they are using.

Finally, declaration practices on *online shopping websites* were analyzed. As shown in Figure 1b and 1c, websites with a web shop do more often declare data collection (in 90% of the cases) then websites without a web shop (with 71%). Three quarter of online shops do inform about cookie usage, but only the half of the websites without shops do so. Similarly, half of the web shops declare the collection of IP addresses, on normal websites only a quart do. However, web analytics tools are seldom mentioned on web platforms, with or without web shops.

3.3 Results on Google Analytics

Google Analytics (GA, 2011a) is the most popular free web analytics tool and is used on more than six million websites (Aden, 2010, Clifton, 2010).

Examining the Forbes 500 companies websites, Google Analytics (GA) was used in 164 cases (33%), followed by Omniture (20%) and Webtrends (16%; compare Figure 2). Smaller software providers like AT Internet, Coremetrics, Nedstat, Unica or eTracker, share the rest of the global web analytics market and were implemented less than in two percent of the websites of the Forbes Global Top 500.

In contrast to Omniture and Webtrends, Google Analytics is relatively less used on communication, service and sales websites. If websites are used for information only, Google Analytics is the market leader (e.g. in the commodity or energy production).

However, on 128 of 164 websites (78%) where Google Analytics was implemented, it was declared that the tool is used to track website traffic. Nevertheless, 22% of the websites using Google Analytics do not declare this. Surprisingly, only 11% of the websites using Google Analytics explicitly declare this, by publishing the official declaration of Google Analytics (in Figure 4). Although Google requires Google Analytics users *to post the official declaration, this is not done in 89% of the cases*.

3.4 Conclusions

Consideration of the first three research questions posed in section 1 leads to the following conclusion:

- Most of the companies **collect web data** and vaguely declare this on the website. Therefore, *most websites meet the minimal requirements* by declaring that data is collected.
- There are considerable **regional differences** in declaring data collection: in *North America* the declaration rate is 91%, however, it is lower in *Europe* (85%) and much lower in *Asia* (70%). Probably for legal reasons, corporate websites from North America declare data collection more often and transparently, although data protection acts and laws are less strict than in Europe.

- Most websites are using **cookies** to identify visitors and visits, but *often do not declare* this at all.
- Most websites do not declare the storage of **IP addresses** and the usage of web analytics **tools**.
- In general, many **declarations** of websites do **not conform** to privacy standards or best practices and could be *much improved*.

4 DECLARATION MODEL FOR WEB DATA COLLECTION

4.1 How to Create Transparency?

The results of the empirical study discussed in section 3 show that the **transparency** regarding data collections on websites and in online shops is low. In addition, there is evidence that many Internet users are uninformed or unaware, of what and how data is collected in the Web, and why. Consequently, there is a need for *sensitization* and clarification for data collection concerns. Since it is difficult on the Internet to internationally regulate data privacy by law, this paper proposes the improvement of *self-regulation* of data collection on websites by website owners and web analytics software providers.

One instrument of self-regulation is **certification** and standardization. Specific certifications or labels of independent third parties (e.g. trust centers) could improve the transparency of the hidden data flow in business practice and insure a certain level of quality regarding best practices and declaration of data collections (Figure 3). Certifications as a *gentleman's agreement* for good web analytics governance would strengthen visitors' *commitment* and *trust* in web analytics, websites and in its different stakeholders.

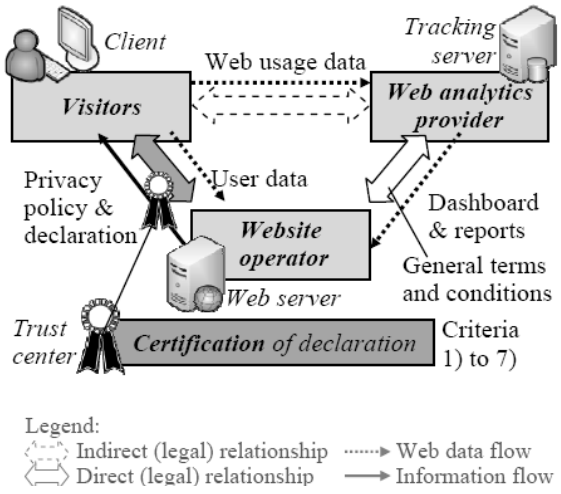

Figure 3: Certification of declaration in web analytics.

4.2 Evaluation & Certification Criteria

The proposed certification leads to the question of *how* ideal declaration is evaluated. The authors suggest the criteria in Table 1 to analyze and rate the level of declaration. The list is neither complete nor exhaustive and needs further discussion.

Table 1: Certification for transparent declaration of web data collection.

Criteria	Description
1) **Declaration** of data collection	Declaration *that* web usage data is collected and stored ("duty to inform" about data collection; under disclaimer, privacy policy or terms and conditions) Specification *what* type of data is collected (e.g. website traffic) Declaration of proprietary *owner* of data
2) **Purpose** of data collection	Declaration *why* data is collected (e.g. for analysis, reporting, optimization of products, services, marketing or website) Declaration of what is happening with the data within the organisation Declaration that only *non-personal data* is stored and for how long If personal data is collected: declaration that *personal data* (e.g. login or address data) is not transferred to third parties Declaration & explicit *permission* of the user, if data is transferred to third parties
3) **Method** of data collection	Declaration of *how* web data is collected, e.g. server-side (logfile analysis), client-side data collection methods (so called page tagging) or the usage of cookies Detailed information, how the data collection method works Further information, *what data* or metrics are collected with this method
4) **Technology** of data collection	Specification of the *technologies* used (e.g. persistent or session cookies, Java Script, Google Analytics Tracking Code) *Description* of the functionalities and handling of technologies, e.g. what cookies are and how they can be deactivated in the web browser Possibility to anonymize *IP addresses* *Localisation* of the tracking server
5) Name of the **software** (provider)	Declaration of the *name* and *version* of the web analytics system used to track website usage and users Information (e.g. name, hyperlink) on the web analytics software *provider*
6) **Opt-out** or opt-in function	Possibility that visitors can *deactivate* data collection during their visits (opt-out function) Alternatively, visitors give their *explicit agreement* to collect data, when the visitors visit the website for the first time (opt-in function) Possibility of *objection* to data collection
7) Right of access and deletion of user/profile data	Declaration that users have the *right of access* to Personally Identifiable Information (PII; personal user/profile data) Possibility to *delete* user data and the history of click behaviour Contact information for privacy issues Publication of a signed "*code of ethics*"

Table 2: Checklist for transparent declaration of data collection on websites.

Certification	Level		Criteria: 1) **Declaration** of data collection	2) **Purpose** of data collection	3) **Method** of data collection	4) **Technology** of data collection	5) **Software** (provider) name	6) **Opt-out**/ opt-in function	7) **Right of access** and deletion of user/profile data
Gold	5	Excellent declaration	✓	✓	✓	✓	✓	✓	✓
Silver	4	Good declaration	✓	✓	✓	✓	✓		
Bronze	3	Adequate declaration	✓	✓	✓				
	2	Minimal declaration	✓	✓					
Legally problematic	1	Insufficient declaration	✓						
	0	No declaration							

4.3 Levels of the Declaration Model

The criteria for transparent declaration of data collection listed in Table 1 can be used in a *declaration model* with different levels shown in Table 2. The higher the level, the better the declaration regarding the transparency and degree of information on data collection on websites.

If a website collects data, but does not declare this in any manner, it is rated as a level 0, *no declaration*. Declaration is *insufficient* (level 1), if information regarding data collection is incomplete, or if the purpose of collection is missing. Level 0 or 1 do not meet national or international data privacy standards and may be legally problematic. The study conducted shows that almost every tenth website is a "black sheep", who does not inform transparently.

However, most websites meet a *minimal or adequate* level 2 and 3 and inform about the collection, purpose and methods of web data collection. To be certified with a “silver medal” for *good declaration* on level 4, websites provide details about technology and software used.

To achieve the “gold medal” for *best practice* in transparent declaration, websites not only have to specify clearly, how, why and to what extent data is collected, but also give users the ability to deactivate data collection and to access or delete user or profile data. Excellent declaration on level 5 confirms with strict data privacy laws of nations like Sweden or of states like Schleswig-Holstein in Germany.

The certification in Table 1 and 2 allows website owners to *evaluate and benchmark* their declarations. If Google Analytics is implemented, the official declaration of Figure 4 has to be published on the website. The model declaration shows that the certification can be applied to analyze or rate any declaration text on the basis of the criteria proposed.

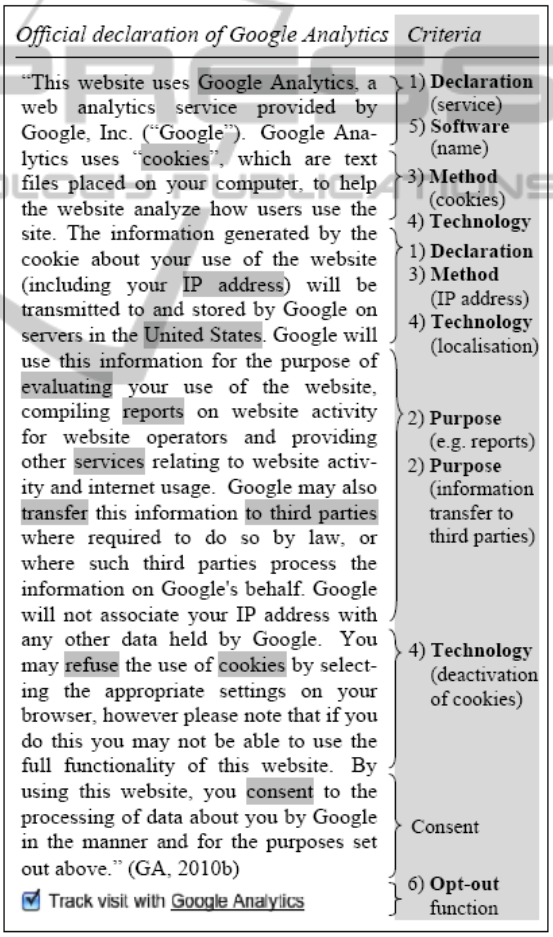

Figure 4: Declaration of Google Analytics (GA, 2011b).

Although the declaration of Google Analytics fulfils many criteria, tool usage is discussed controversially, since Google stores data in the US, connects it with other data, does not anonymize IP addresses and gives data to third parties but no right of access.

5 CONCLUSIONS

5.1 Recommendations

Recommendations to Internet Users and Visitors

- Be aware that your *every click is tracked* on most websites and web shops by client- and/or server-side data collection methods, and analyzed by software, even though this is not declared.
- Install *WASP* in your Mozilla Firefox browser if you want to know which client-side systems are tracking you on your trip through the cyberspace.
- Install the Google Analytics deactivation-add-on, if you do not want to be tracked by *Google*.
- Alternatively, you can deactivate *JavaScript* in your browser settings, if you do not want to be tracked by any client-side web analytics system.
- Websites indentify your computer by *cookies* or the *IP address*, which is often undeclared.
- *Deactivate cookies* in your web browser or delete them regularly if you don't want to be identified.

Recommendations to Data Protection Officers

- Be aware that many websites, even of big enterprises, do not meet international *privacy standards and laws* of data collection and declaration.
- *Inform* and educate the public about web analytics concerns and about the current privacy policy problems on websites and especially in e-shops.
- Control and improve websites compliances with privacy policies and *data protection acts* like the "Directive 95/46/EC of the European Parliament and of the Council" (EUR-Lex, 2011).
- Support local, national and international recommendations, guidelines, labels and *certifications* that guarantee a legal and transparent declaration of data collections on websites.

Recommendations to Webmasters and Managers

- *Declare* data collection transparently, honestly and forthrightly. Ensure that privacy policies fulfil data protection acts and are up-to-date.
- Hold data privacy in your highest regard and do everything to keep data *secure and private*.
- Do not link web analytics data to profile or *personal* data like names, addresses, phone numbers or payment information like credit card numbers.
- If you are using *cookies*, this information should be included in the privacy policy.
- If you track IP addresses, you should declare this. If possible, make *IP addresses anonymous*.
- The declaration of the data collected is necessary to *create confidence* and to sustain credibility or reputation of your website and your organization.

5.2 Critical Discussion & Outlook

This study analyzed, which client-side web analytics systems are used to gather web data on websites of international companies and if they declare data collection. However, companies have not been contacted directly to interview them regarding their web data collection. Therefore, we do not know, *what* they do exactly with the collected web data. We only know that most websites do not inform properly about their collecting practices. In addition only websites of big companies were reviewed, but no sites of small and medium-sized companies. Probably, SMEs declare even less about data collections. An independent web analytics certification platform would help to improve general declaration practices.

So far, *little practical and academic research* has been done on web analytics systems. However, as Internet traffic increases, this research field will gain in importance. Other studies in web analytics and data privacy are absolutely essential to better understand and manage the nature, *chances and risks* of the dynamic and fast-developing web environment.

REFERENCES

Aden, T. (2010). *Google Analytics*, Hanser, München.

Clifton, B., 2010. *Advanced Web Metrics with Google Analytics*, Wiley. New York.

EUR-Lex (2011). *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data*, Retrieved on February 6, from: http://eur-lex.europa.eu/.

Kaushik, A. (2009). *Web Analytics 2.0* Wiley. New York.

Forrester Consulting (2009). *Appraising Your Investments In Enterprise Web Analytics*, Retrieved on February 6, 2011, from: http://www.forrester.com.

Forbes (2011). *Forbes: The Global 2000*, Retrieved on February 6, from: http://www.forbes.com.

GA (2010a). *Google Analytics*, Retrieved on February 6, from: http://www.google.com/analytics.

GA (2010b). *Google Analytics Terms of Service*, Retrieved on February 6, from: http://www.google.com /intl/en_uk/analytics/tos.html.
Meier, A., Stormer, H. (2009). *eBusiness and eCommerce: Managing the Digital Value Chain*, Springer, Berlin.
Meier, A. Zumstein, D. (2010). *Web Analytics – An Introduction (in German)*, dpunkt, Heidelberg.
Sterne, J., 2002. *Web Metrics*, Wiley. New York.
WAA (2011). *Web Analytics Association*. Retrieved on February 6, from: www.webanalyticsassociation.org.
WASP (2011). Retrieved on February 6, from: http://web analyticssolutionprofiler.com.
Zumstein, D., 2010. In: *6th Int. Conf. on Web Information Systems & Technologies*, April 7-10, Valencia, Spain.